Question 1

What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

Answer:

The optimal values of alpha for Ridge regression is 0.3 and for Lasso Regression is 10.

There are no change in predictor variables when the alpha values are doubled. However after doubling the alpha it was observed that the model coeff values have been reduced due to higher cost function penalty.

Top 5 Model Coefficients are

	Ridge2	Ridge	Lasso	Lasso2
GrLivArea	119647.820963	121129.038711	1.892914e+05	2.061966e+05
1stFlrSF	100871.422079	102414.710407	4.300321e+04	2.753867e+04
BsmtFinSF1	87197.397102	87052.075608	8.587464e+04	7.944806e+04
OverallQual_VP	64285.977000	67582.253605	6.993505e+04	4.268879e+04
GarageArea	59112.585887	59915.927764	5.996781e+04	5.098918e+04

Ridge2 and Lasso2 model coefficients for which alpha value was doubled.

Sight improvement in R2 score of the model was also observed due to elimination of effects of overfitting.

Question 2

You have determined the optimal value of lambda for ridge and lasso regression during the assignment. Now, which one will you choose to apply and why?

Answer:

Below are given are performance parameters for both models on train and test data.

	Metric	Ridge Regression	Lasso Regression
0	R2 Score (Train)	8.674315e-01	8.675908e-01
1	R2 core (Test)	8.770256e-01	8.766068e-01
2	RSS (Train)	6.820391e+11	6.812199e+11
3	RSS (Test)	2.690416e+11	2.699579e+11
4	MSE (Train)	2.731691e+04	2.730050e+04
5	MSE (Test)	2.616456e+04	2.620908e+04

It is observed that the R2 score on test data is slighltly better for Ridge Regression on test data.

However it is also observed that Lasso Regression uses fewer number of predictor variables for arriving at the similar R2score.

	Ridge2	Ridge	Lasso	Lasso2
GrLivArea	119647.820963	121129.038711	1.892914e+05	2.061966e+05
BsmtFinSF1	87197.397102	87052.075608	8.587464e+04	7.944806e+04
OverallQual_VP	64285.977000	67582.253605	6.993505e+04	4.268879e+04
RoofMatl_Metal	42594.160739	57137.378192	6.691381e+04	0.000000e+00
GarageArea	59112.585887	59915.927764	5.996781e+04	5.098918e+04
OverallQual_P	56256.519408	55644.581493	5.529836e+04	6.091141e+04
GarageYrBlt_2009.0	44811.669273	44923.449291	4.480936e+04	3.985386e+04
1stFlrSF	100871.422079	102414.710407	4.300321e+04	2.753867e+04
MSZoning_RL	35931.220916	41086.495994	4.203545e+04	7.644126e+03
BsmtUnfSF	38964.918841	38607.640375	3.716744e+04	2.957521e+04
MSZoning_RH	28018.435532	34546.558245	3.585694e+04	0.000000e+00
MSZoning_FV	28765.755391	34392.318578	3.564732e+04	0.000000e+00
MSZoning_RM	28767.847995	34301.638938	3.547509e+04	-0.000000e+00
YearBuilt	33751.524476	33500.855666	3.349169e+04	3.591315e+04
2ndFlrSF	48293.244520	48388.723562	1.202229e+04	0.000000e+00
Exterior1st_CemntBd	4491.980107	4004.870255	7.083460e+03	9.839359e+03
HouseStyle_2.5Unf	4271.570951	4891.455161	6.445038e+03	0.000000e+00
GarageCond_Gd	3721.256635	2666.498022	6.374278e+03	0.000000e+00
OverallQual_VE	4584.758018	8086.197134	5.382792e+03	-0.000000e+00
MSSubClass_75	4271.570951	4891.455161	2.321228e+03	0.000000e+00
GarageQual_Gd	2600.355451	2235.144652	5.362362e+02	0.000000e+00

Since the Lasso Regression model is more simpler and as accurate as the Ridge Regression model, we will go for Lasso Regression Model as our final model.

Question 3

After building the model, you realised that the five most important predictor variables in the lasso model are not available in the incoming data. You will now have to create another model excluding the five most important predictor variables. Which are the five most important predictor variables now?

Answer:

The final model selected is using Lasso

The new Lasso Regression model coefficient after removing top 5 predictor variables are.

	Lasso
1stFlrSF	3.264126e+05
2ndFlrSF	1.159057e+05
RoofMatl_Metal	6.434582e+04
YearBuilt	6.042583e+04
GarageYrBlt_2009.0	5.476292e+04

Question 4

How can you make sure that a model is robust and generalisable? What are the implications of the same for the accuracy of the model and why?

Answer:

The R2 score for the test and training set should be almost equal. A models predictive power on the test data shows how good a model is.

In order to ensure that the model has good predictive power on the test data, it has to be ensured that there is no overfitting. We use Ridge regression and Lasso Regression for elminating the possibilty of overfitting.

Ridge and Lasso includes a cost function penalty for the every predictor variable. The penalty value depends directly with the value of model coefficients the predictor variable caries.

Ridge and Lasso works by sacrificing small bias for significant reduction in variance of the model.

The implications of using a Ridge and Lasso regression is that the model arrived is not overfitting the training data. This means the when the model will have proper predictive power when used on the test data as implied by the very close R2 Score when the model is applied on train and test data.